Research Statement Ryan Giordano

Many researchers would be concerned if they learned that some core conclusion of their statistical analysis—such as the sign or statistical significance of some key effect—could be overturned by removing a small fraction of their data. Such non-robustness would be particularly concerning if the data were not actually drawn randomly from precisely the population of interest, or if the model may have been misspecified—circumstances that often occur in the social sciences. For example, a recent study of microcredit in Mexico on 16,561 housholds measured a negative but statistically insignificant effect of microcredit [Angelucci et al., 2015]. However, in recent work, my co-authors and I show that one can make the estimated effect of microcredit positive and statistically significant by removing just 15 households from the analysis, reversing the paper's qualitative conclusions by removing only 0.1% of the data. Since there are a combinatorially large number of ways to leave 15 datapoints out of 16,561 (over 10^{51}), finding such influential subsets by brute force would be impossible. I circumvent this difficulty by forming a *linear approximation* to the dependence of the estimator on the dataset. My technique works for a wide class of commonly used estimators, providing a fast, automatic tool for identifying small but influential subsets with finite-sample error bounds in many applied problems.

In general, my research provides fast, easy-to-use computational tools for assessing whether small changes to modeling assumptions or datasets can meaningfully alter an analyst's substantive conclusions. As I detail below, my work encompasses Bayesian approaches (such as measuring robustness to prior specification), frequentist approaches (such as fast approximations to leave-one-out cross validation), and combinations of the two (such as measuring the robustness of Markov chain Monte Carlo procedures to data resampling). For the remainder of this statement, I will describe some of my specific past and future projects, emphasizing the ways in which I update classical results for a modern computational environment using intuitive, relevant theory, and with motivation drawn strongly from practical applications.

1 Selected prior work

1.1 Robustness to data ablation

In many applied settings, particularly in econometrics, a statistical analysis might be considered non-robust if it could be overturned or even reversed by removing only a small proportion of the dataset. Analyzing all possible data subsets of a certain size is typically computationally prohibitive, so I provide a finite-sample metric to approximately compute the number (or fraction) of observations that has the greatest influence on a given result when dropped [Broderick et al., 2020][1]. At minimal computational cost, my method provides an exact finite-sample lower bound on sensitivity, so any non-robustness one finds is conclusive. I demonstrate that non-robustness to data ablation is driven by a low signal-to-noise ratio in the inference problem, is not reflected in standard errors, does not disappear asymptotically, and is not inherently a product of outliers or misspecification.

The approximation works for M-estimators based on smooth estimating equations, a class which includes most standard estimators, including ordinary least squares, instrumental variables, generalized method of moments, variational Bayes, and maximum likelihood estimators. Using my `R` package [Giordano, 2020], the approximation is automatically computable from the specification of the estimating equation alone. By analyzing several published econometric analyses, I show that even two-parameter linear regression analyses of randomized trials can be highly sensitive. While I find some applications are robust, in others the sign of a treatment effect can be changed by dropping less than 1% of the sample even when standard errors are small.

[1]Following conventions in econometrics, the authors are listed alphabetically. Rachael Meager and I are equal contribution primary authors.

1.2 Approximate cross validation

The error or variability of machine learning algorithms is often assessed by repeatedly re-fitting a model with different weighted versions of the observed data; cross-validation (CV) can be thought of as a particularly popular example of this technique. In Giordano et al. [2019b], I use a linear approximation to the dependence of the fitting procedure on the weights, producing results that can be faster by an order of magnitude than repeated re-fitting. I provide explicit finite-sample error bounds for the approximation in terms of a small number of simple, verifiable assumptions. My results apply whether the weights and data are stochastic or deterministic, and so can be used as a tool for proving the accuracy of the approximation on a wide variety of problems. As a corollary, I state mild regularity conditions under which the approximation consistently estimates true leave-k-out cross-validation for any fixed k. I demonstrate the accuracy of the approximation on a range of simulated and real datasets, including an unsupervised clustering problem from genomics.

1.3 Prior sensitivity for Markov chain Monte Carlo

Prior specification encodes key assumptions in Bayesian statistics, but Bayesian inference can be sensitive to prior specification. Evaluating the sensitivity of Bayesian posterior expectations to prior specification by re-estimating the posterior for a large number of alternative priors is typically computationally prohibitive. In particular, Markov chain Monte Carlo (MCMC) is arguably the most commonly used computational tool to estimate Bayesian posteriors, which is made still easier by modern black-box MCMC tools such as `Stan`. However, a single run of MCMC typically remains time-consuming, and systematically exploring alternative prior parameterizations by re-running MCMC would be computationally prohibitive for all but the simplest models.

My software package, `rstansensitivity`, [Giordano, 2018, Giordano et al., 2018b], takes advantage of the automatic differentiation capacities of `Stan` together with a classical result from Bayesian robustness [Gustafson, 1996, Giordano et al., 2018a] to provide automatic hyperparameter sensitivity for generic `Stan` models from only a single MCMC run. I demonstrate the speed and utility of the package in detecting excess prior sensitivity in a social sciences model taken from Gelman and Hill [2006, Chapter 13.5].

1.4 Approximately bootstrapping Bayesian posterior means

When analyzing randomly sampled data using Bayesian posteriors, it can make sense to consider the *sampling variability* of the posterior mean. For example, one might ask whether a new random sample of poll respondents in the presidential forecast model of Gelman and Heidemanns [2020] would lead to a different prediction for the election outcome. When Bayesian models are misspecified, or when they contain latent parameters on which one wishes to condition when resampling, the sampling and posterior variances can differ. In such cases, sampling variability in excess of posterior variability is symptomatic of *data non-robustness*—new data sets which are *a priori* plausible would lead to different substantive conclusions.

The sampling variability of a posterior mean can be evaluated by the bootstrap, but at the considerable cost of re-running Markov chain Monte Carlo (MCMC) hundreds of times. In Giordano and Broderick [2020, 2021], I propose an efficient alternative to bootstrapping an MCMC procedure. My approach is based on the Bayesian analogue of the influence function from the classical frequentist robustness literature. Using results from Giordano et al. [2018a, 2019b], I show that the influence function for posterior expectations can be easily computed from the posterior samples of a single MCMC procedure and consistently estimates the bootstrap variance. I demonstrate the accuracy and computational benefits of the influence function variance estimates on an array of experiments including an election forecasting model, the Cormack-Jolly-Seber model from ecology, and a large collection of models and datasets from the social sciences.

1.5 Prior sensitivity for discrete Bayesian nonparametrics

A central question in many probabilistic clustering problems is how many distinct clusters are present in a particular dataset and which observations cluster together. Discrete Bayesian nonparametric (BNP) mixture models address this question by placing a generative process on cluster assignment, making the number and composition of distinct clusters present amenable to Bayesian inference. However, like all Bayesian approaches, BNP requires the specification of a prior, and this prior may favor a greater or lesser number of distinct clusters.

In Giordano et al. [2021], I derive and analyze prior sensitivity measures for variational Bayes (VB) approximations in general, with a practical focus on discrete BNP models. Unlike much previous work on local Bayesian sensitivity for BNP (e.g. Basu [2000]), I pay special attention to the ability of the sensitivity measures to *extrapolate* to different priors, rather than treating the sensitivity as a measure of robustness *per se*. Under mild regularity conditions, I prove that VB approximations are Fréchet differentiable functions of the prior density, though only in one extreme of a standard family of embeddings considered for exact Bayesian posteriors [Gustafson, 1996]. My co-authors and I apply the sensitivity measures to a number of real-world problems, including an unsupervised clustering problem from genomics using fastSTRUCTURE, demonstrating that the approximation is accurate, orders of magnitude faster than re-fitting, and capable of detecting meaningful prior sensitivity in quantities of practical interest.

1.6 Uncertainty propagation in mean-field variational Bayes

Mean-field Variational Bayes (MFVB) is an approximate Bayesian posterior inference technique that is increasingly popular due to its fast runtimes on large-scale scientific data sets (e.g., Regier et al. [2019]). However, even when MFVB provides accurate posterior means for certain parameters, it often mis-estimates variances and covariances due to its inability to propagate Bayesian uncertainty between statistical parameters.

In Giordano et al. [2015, 2018a], I derive a simple formula for the effect of infinitesimal perturbations on MFVB posterior means, thus providing improved covariance estimates and greatly expanding the practical usefulness of MFVB posterior approximations. My method for computing posterior covariances from an MFVB approximation exploits a result from the classical Bayesian robustness literature relating derivatives of posterior expectations to posterior covariances, and can be seen as generalizing the Laplace approximation for maximum *a-posteriori* estimates to more general MFVB procedures. In experiments on simulated and real-life datasets, including models from ecology, the social sciences, and on a massive internet advertising dataset, I demonstrate that my method is simple, general, and fast, providing accurate posterior uncertainty estimates and robustness measures with runtimes that can be an order of magnitude faster than MCMC.

2 Selected future work

Though there are many potential directions for my future research, I articulate below two broad areas that I find particularly exciting.

2.1 The empirical influence function

Much of my work (particularly Giordano et al. [2019b], Broderick et al. [2020], Giordano and Broderick [2021]) has strong connections to the classical theory of von Mises expansions and the closely related concept of the influence function, which measures the effect of individual datapoints on an estimator [Mises, 1947, Reeds, 1976, Hampel, 1986, Serfling, 2009]. But my focus on the influence function evaluated at the observed data—i.e., the "empirical influence function" (EIF)—stands in contrast with much of the classical literature, which studies the asymptotic behavior of estimators via their (unobserved) limiting influence function. In our present age of automatic differentiation, large datasets, and complex models, I believe that the EIF will continue to provide practical benefits and is relatively under-studied.

In Giordano et al. [2019a], I show that higher-order EIFs can be easily and automatically evaluated and analyzed for M-estimators at a computational cost comparable to the first-order EIF—that of forming and factorizing a Hessian matrix of second-order derivatives. Thus, the EIF "amortizes" the cost of evaluating an M-estimator large number of alternative datasets: by paying a large fixed price up front (approximately computing and factorizing a Hessian matrix), one can cheaply approximate M-estimators at a very large number of alternative datasets. Natural applications of the idea include approximating the bootstrap-after-bootstrap, evaluating the sampling properties of cross-validation, and computing higher-order jackknife bias correction.

2.2 Sensitivity analysis in difficult situations

It is not always as easy to apply sensitivity analysis in practice as it is in theory. I have found that a few key problems tend to recur, and I will discuss them in turn, as well as potential solutions which draw connections to the optimization literature.

First, sensitivity analysis should, ideally, deal gracefully with incomplete optimization. For example, a collaborator from biostatistics and I have found that the `R` package `DESeq2` package can fail in practice to fully optimize the log likelihood, and so fail to satisfy the assumptions that make sensitivity analysis possible. Instead of forcing users to optimize further, I propose that second-order empirical influence functions could simulate the effect of simultaneously taking a Newton step and perturbing the data, permitting sensitivity analysis on incompletely optimized objectives with little computation beyond that required for well-optimized objectives.

Second, the key computational bottleneck in sensitivity analysis in high-dimensional problems is typically the solution of linear systems involving the inverse Hessian of the objective function. Off-the-shelf iterative algorithms like the conjugate-gradient algorithm suffice in many cases, but there is reason to believe that the present active research into stochastic second-order methods could significantly speed up sensitivity analysis in large problems.

Finally, practitioners are often interested in non-smooth objectives. For these, one promising idea is to use local approximations to speed up computationally intensive but smooth components in non-smooth problems. For example, Wilson et al. [2020] speeds up cross-validation of linear regression with a non-smooth lasso penalty by forming a fast approximation to the effect on the optimal squared error of leaving out a single datapoint, and retaining non-smoothness in the lasso penalty. We take a similar approach in Giordano et al. [2021], retaining easy-to-compute non-linearities in posterior summary statistics.

2.3 Projects for junior collaborators

Finally, a professor must also be able to provide research topics suitable for more junior researchers, such as those who are early in their PhD. To this end, many of my existing papers can be easily "crossed" with one another, producing relatively straightforward projects for more junior collaborators. For example, I am presently working with a PhD student to combine the Bayesian influence function of Giordano and Broderick [2021] with the adversarial data ablation metrics of Broderick et al. [2020], to automatically find influential data subsets from Markov chain Monte Carlo output.

References

Angelucci, M., Karlan, D., and Zinman, J. (2015). Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82.

Basu, S. (2000). Bayesian robustness and Bayesian nonparametrics. In Insua, D. R. and Ruggeri, F., editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media.

Broderick, T., Giordano, R., and Meager, R. (2020). An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *arXiv preprint arXiv:2011.14999*. Note: Following conventions in econometrics, the authors are listed alphabetically. Giordano and Meager are equal contribution primary authors.

Gelman, A. and Heidemanns, M. (2020). The Economist: Forecasting the US elections. Data and model accessed Oct., 2020.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Giordano, R. (2018). StanSensitivity: Automated hyperparameter sensitivity for Stan models. GitHub repository `https://github.com/rgiordan/StanSensitivity`.

Giordano, R. (2020). Zaminfluence. GitHub repository `https://github.com/rgiordan/zaminfluence`.

Giordano, R. and Broderick, T. (2020). Effortless frequentist covariances of posterior expectations in Stan. Presentation at Stancon 2020 `https://tinyurl.com/y2e2ucp3`.

Giordano, R. and Broderick, T. (2021). The Bayesian infinitesimal jackknife for variance. *In preparation.*

Giordano, R., Broderick, T., and Jordan, M. (2018a). Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029.

Giordano, R., Broderick, T., and Jordan, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449.

Giordano, R., Broderick, T., and Jordan, M. I. (2018b). Automatic robustness measures in Stan. Presentation at Stancon 2018 `https://tinyurl.com/yyqwpowc`.

Giordano, R., Jordan, M. I., and Broderick, T. (2019a). A higher-order Swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116.*

Giordano, R., Liu, R., Jordan, M. I., and Broderick, T. (2021). Evaluating sensitivity to the stick breaking prior in Bayesian nonparametrics. *https://arxiv.org/abs/2107.03584*.

Giordano, R., Stephenson, W., Liu, R., Jordan, M. I., and Broderick, T. (2019b). A Swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147.

Gustafson, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195.

Hampel, F. (1986). *Robust statistics: The approach based on influence functions*, volume 196. Wiley-Interscience.

Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348.

Reeds, J. (1976). *On the definition of von Mises functionals*. PhD thesis, Ph. D. Thesis, Statistics, Harvard University.

Regier, J., Fischer, K., Pamnany, K., Noack, A., Revels, J., Lam, M., Howard, S., Giordano, R., Schlegel, D., and McAuliffe, J. (2019). Cataloging the visible universe through Bayesian inference in Julia at petascale. *Journal of Parallel and Distributed Computing*, 127:89–104.

Serfling, R. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Wilson, A., Kasy, M., and Mackey, L. (2020). Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*, pages 4530–4540. PMLR.